GHAJAR
EXHIBIT 23

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN FRANCISCO DIVISION

RICHARD KADREY, et al., Individual and Representative Plaintiffs, v. META PLATFORMS, INC., a Delaware corporation; Defendant.	Case No. 3:23-cv-03417-VC

HIGHLY CONFIDENTIAL - ATTORNEYS' EYES ONLY
VIDEO-RECORDED DEPOSITION OF CRISTINA LOPES, PH.D.

Thursday, February 13, 2025
San Francisco, California

Stenographically Reported By:
Hanna Kim, CLR, CSR No. 13083
Job No. 7173634

THE COURT REPORTER: "Predict the next"?

THE WITNESS: -- the next token. So it's -- it's con- -- the -- during training, it's configured so that the inputs are the sequences of text. The outputs are the same sequences of text, but shifted by one, which gives the next over. 10:03:44

And so the -- the model goes through several days, maybe weeks of doing that training and -- until it converges to some acceptable behavior. 10:04:05

BY MR. WEINSTEIN:

Q. At the end of a pretraining process, is the result of that something called a base model?

A. That's how it's commonly called, yes.

Q. Okay. And a base model is a model that has been pretrained but doesn't have any fine-tuning applied to it; correct? 10:04:15

A. Correct.

Q. Okay. And in order to pretrain a large language model, does that require a large amount of data? 10:04:31

A. Yes.

Q. An enormous amount of data?

A. I mean, it depends on what you want to do with the large language model. So there's -- 10:04:44

large -- you know, large language models can also be for specific niche applications, like, for example, code or software engineering tasks.

So it -- it depends if you -- you -- on the other hand, if you want to train a large language model to be multitasking, multifunctionality, you know, do everything, then you -- you -- you need a lot of data. 10:04:55

Q. All right.

Had you ever worked with -- yeah, withdrawn. 10:05:05

Prior to your work on this case, had you ever worked with any of the Meta large language models?

A. No. 10:05:13

Q. Okay. But obviously since your work in this case you've been working with some of the Llama models; correct?

A. Yes.

Q. Okay. And you -- you have actually used the Llama 3 7B and 70B models; correct? 10:05:27

A. Yes.

Q. Okay.

A. And 8B.

Q. 8B. 10:05:36

Q. Understood.

And that -- that's not just limited to open-source data, that's any kind of text data --

A. Any --

Q. -- correct? 10:06:58

A. -- any kind of text data.

THE COURT REPORTER: We need to speak one at a time, please.

THE WITNESS: I'm sorry.

MR. WEINSTEIN: I violated my own rule on that. 10:07:06

BY MR. WEINSTEIN:

Q. Have you ever heard the term "overfitting" as it refers to large language models?

A. Yes. 10:07:16

Q. What is "overfitting"?

A. It's when the network -- when the neural network starts to learn too much about the data and starts just basically memorizing the data and -- and loses its ability to generalize from it. 10:07:30

Q. Is deduplication of training data an example of a technique that helps prevent overfitting?

A. Yes.

Q. Okay. What other techniques that -- 10:07:43

outputs and the obtained outputs. When that happens, training can end, and the model is deemed ready for use or 'inference.'" [As read]

That's correct; right?

A. Correct. 10:25:53

Q. And so just so we're clear, the term "inference," that's a term that you use to describe using the model after it's been at least in the pretraining stage; correct?

A. Correct. 10:26:03

Q. Okay. And then Paragraph 40, you refer to "generalization." And you say in the middle of the Paragraph, "The value of deep learning models comes from the value of mathematical regression itself: the ability to produce reasonable outputs even when 10:26:27 the inputs were not in the training data. This is called 'generalization.'"

Do you see that?

A. Yes.

Q. Is generalization a desired characteristic 10:26:33 of a large language model?

A. Yes.

Q. And people who build large language models, they strive to make their models generalized as much as possible; correct? 10:26:55

MR. YOUNG: Objection to form.

THE WITNESS: I believe so.

BY MR. WEINSTEIN:

Q. Okay. And by generalization, we're referring to the ability of the model to provide output that is different from the data on which it was trained; correct? 10:27:12

A. It's -- it refers to the ability of the model to produce reasonable outputs on data on which it was not trained. 10:27:31

Q. Right.

Obviously, it's -- it's not any random data, you want reasonable outputs; right?

A. Yes.

Q. The -- the goal is to provide an output that was not in the original training data; correct? 10:27:39

A. Correct.

Q. Right.

You also say that, in Paragraph 40, "Generalization stands on the quality of the training data." 10:27:51

Do you see that?

A. Yes.

Q. When you talk about training data quality, that's not necessarily the same thing as the quality 10:28:00

BY MR. WEINSTEIN:

Q. So by "high quality," is this referring to the property of the data, that it's expected to cause an improvement in the model's performance?

A. Yes. 10:30:43

Q. Okay. And then you mentioned Wikipedia.

And I realize you may not have researched this, but what is the reason that Wikipedia is regarded as high-quality training data, assuming it is? 10:31:01

MR. YOUNG: Objection. Form.

Go ahead.

THE WITNESS: I can -- I -- I -- I don't know. I can give an educated guess. It's because it's highly curated. It has a lot of eyes on it. 10:31:10

The text tends to be very good. The explanations tend to be very factual. There's -- references back up the statements in most of the articles.

So I would imagine that that's the reason why the LLM developers like Wikipedia so much. 10:31:30

BY MR. WEINSTEIN:

Q. And as -- as far as books you mentioned, why, in your opinion, do books provide a high quality of training data?

A. Again, I am trying -- I -- I -- I have -- 10:31:42

Common Crawl has 61 trillion tokens.

Do you see that?

A. Yes.

Q. You'd agree that not all the tokens in Common Crawl are human readable text; correct? 10:45:36

A. I don't know.

Q. Well, for example, there may be HTML elements in the web pages that are crawled [verbatim]; correct?

A. I would imagine, yes. 10:45:50

Q. There may be nontext media elements; correct?

A. Probably, yes.

Q. URLs?

A. Yes. 10:45:58

Q. Metadata?

A. Yes.

Q. Okay. You reference quality as being an important consideration in training data.

Would you agree that's not the only 10:46:18 important consideration in terms of training data used for an LLM?

A. Yes. So -- yeah.

Q. You also need a lot of training data; correct? 10:46:33

A. That's correct.
Q. And you need a wide diversity of training data; correct?
A. So, as I said, it depends on how you want to use your large language model for; right? 10:46:44
So if you're talking about a -- a large language model that wants to do everything, then you want a lot of diversity. You want to be able to capture everything.
Q. And Llama would be an example of a large 10:46:54 language model that would want to do everything?
MR. YOUNG: Objection to form.
THE WITNESS: I believe so.
BY MR. WEINSTEIN:
Q. The same would probably apply to ChatGPT; 10:47:06 correct?
A. Yeah.
MR. YOUNG: Object to form.
BY MR. WEINSTEIN:
Q. If we could turn to Paragraph 87, there's 10:47:15 a reference to "Meta downloaded a dataset known as 'The Pile.'" [As read]
Do you see that?
A. Mm-hmm. Yes.
Q. Now, I'll represent to you there's another 10:47:36

Q. Possibly thousands of times a day; is that correct?

MR. YOUNG: Objection to form.

THE WITNESS: I don't know the number.

BY MR. WEINSTEIN: 02:17:30

Q. Okay. And so, would you agree with him that the likely reason that his tests showed a higher degree of memorization for Embraced was because of the presence of Bible passages that were otherwise available on the internet? 02:17:47

A. It's a -- it's a possible explanation, yes.

Q. Okay. Why don't we turn to Paragraph 91. And this is a looking at some tests that you ran using Llama 3 70B; is that right? 02:18:37

A. Yes.

Q. And these were tests that were run on a server at UCI?

A. Yes.

Q. Okay. Using a base model; correct? 02:18:48

A. Correct.

Q. Okay. And in this experiment, you said that "Counsel provided me with 73 passage drawn from a variety of books, including Asserted Works, to be used as prompts for a completion using the Llama 8B 02:19:06

and the Llama 3 70B models." [As read]

Do you see that?

A. Yes.

Q. So your tests were using the same technique of providing text from a book and seeing if the model could complete it; correct? 02:19:16

A. Correct.

Q. Okay. So the 73 passages that you were provided, they were selected by counsel for you?

A. Yes. 02:19:34

Q. Okay. And in a footnote, you're citing to a -- a couple of LinkedIn posts, it looked like from -- from the initials, Louis W. Hunt?

A. Yes.

Q. Have you ever spoken to Louis -- 02:19:54

A. No.

Q. -- Hunt?

A. No.

Q. Do you know who he is?

A. No. 02:20:00

Q. Okay. The samples that you were provided -- withdrawn.

The passages that you were provided, were they passages that were from Mr. Hunt's work?

A. Yes, I believe so. 02:20:18

Q. Okay. So Mr. Hunt published a series of examples of passages and showed a certain degree of completion for some of the passages using Llama; is that right?

A. That's right. 02:20:40

Q. Okay. And the footnotes show in one case, he had 400 pages of algorithmically generated activity. The other 131, did you look at any of the other ones that he did other than the 73 that counsel provided to you? 02:21:01

A. No.

Q. Just the 73, okay.

A. Yes.

Q. What was the -- see if there's a way I can ask this without getting into Rule 26 territory. 02:21:22

What was the basis for the selection of the 73 passages to which you were provided?

MR. YOUNG: I'm going to caution the witness to not reveal any discussions she may have had with counsel or any -- reveal any mental 02:21:35 impressions she -- of counsel.

But to the extent you are able to answer, please do.

THE WITNESS: Okay. Can you repeat the question. 02:21:48

a large language model to be able to repeat part of input text is a function of how many times the input text was present in the dataset; correct?

A. Yes, I understand that.

Q. So would it be your expectation that when Llama is able to complete certain passages, that those passages likely appear multiple times in the dataset? 03:32:44

A. Yes. That's why de-duplication is so important. 03:33:02

Q. Now, all these tests that were done to confirm the ability to complete passages, they all involved inputting text from the original work as the prompt; correct?

A. Yes. 03:33:19

Q. So to -- in -- in order to run these tests, somebody who created these prompts would have had to have access to the original text; correct?

A. Yes.

MR. YOUNG: Objection to form. 03:33:31

BY MR. WEINSTEIN:

Q. Have you heard the term "adversarial prompt" before?

A. Yes.

Q. What's an adversarial prompt? 03:33:40

A. These are prompts that people use to try to coerce these large language models to misbehave in some way.

Q. Understood.

The prompts that were identified for purposes of these memorization studies in Appendix C of your report, are those examples of adversarial prompts? 03:33:53

A. I don't think so. I think these are just typical examples of, you know, trying to complete things. Now, the -- the collection -- if you ask me about the collection, like who collected these prompts, it was not me. These prompts were given to me. I -- I don't -- you know, I don't know why, when, what went into the selection of these prompts, as I said before. 03:34:07 03:34:29

Q. Right. But the per- -- person who put -- put together the prompts was trying to get the model to output text from these books; correct?

MR. YOUNG: Objection to form and scope. 03:34:41

THE WITNESS: I -- I think that's what he was trying to do. I mean, but that's also what all of these memorization experiments are trying to do; right?

BY MR. WEINSTEIN: 03:34:52

Q. Right. But this type of use of a large language model to complete passages with original text, is that a use you've done in -- is that part of ordinary use of a large language model?

A. No. This is part of checking the -- you know, the soundness and the -- the behavior of large language models is part of something that -- you know, like red teaming does, for example. So this -- 03:35:09

THE COURT REPORTER: Was the what? 03:35:19

THE WITNESS: Red teaming. So groups that are, even inside the same development, the same company or maybe in other companies that try to really poke at the system and make it mis- -- misbehave so that, you know, measures and mitigations can be taken. 03:35:38

BY MR. WEINSTEIN:

Q. And -- and all these tests were run on the base model; correct?

A. Yes. 03:35:47

Q. So if there were mitigations in place to try to prevent output that repeats training data as part of post training, those wouldn't have been effective in these experiments; correct?

MR. YOUNG: Form and scope. 03:36:02

CERTIFICATE OF REPORTER

I, Hanna Kim, a Certified Shorthand Reporter, do hereby certify:

That prior to being examined, the witness in the foregoing proceedings was by me duly sworn to testify to the truth, the whole truth, and nothing but the truth;

That said proceedings were taken before me at the time and place therein set forth and were taken down by me in shorthand and thereafter transcribed into typewriting under my direction and supervision;

I further certify that I am neither counsel for, nor related to, any party to said proceedings, not in anywise interested in the outcome thereof.

Further, that if the foregoing pertains to the original transcript of a deposition in a federal case, before completion of the proceedings, review of the transcript [X] was [ ] was not requested.

In witness whereof, I have hereunto subscribed my name.

Dated: February 28, 2025.

Hanna Kim, CLR, CSR No. 13083